PROJECT NASTY- NEWS AGGREGATION AND SUMMARIZATION

A

MAJOR PROJECT REPORT

Submitted by
SUPRATIK CHATTERJEE (RA1511003040196)

Under the guidance of

Mr. Muthurasu N., B.E., M.Tech.,

(Assistant professor, Department of Computer Science Engineering)

in partial fulfilment of the degree
Of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

SRM INSTITUTE OF TECHNOLOGY AND SCIENCE
Vadapalani, Chennai-600026

APRIL 2019

PROJECT NASTY- NEWS AGGREGATION AND SUMMARIZATION

A

MAJOR PROJECT REPORT

Submitted by
SUPRATIK CHATTERJEE (RA1511003040196)

Under the guidance of

Mr. Muthurasu N., B.E., M.Tech.,

(Assistant professor, Department of Computer Science Engineering)

in partial fulfilment of the degree
Of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF TECHNOLOGY AND SCIENCE
Vadapalani, Chennai-600026

APRIL 2019

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

It is certified that this project report titled "Project NASTY" is the bonafide work of **Supratik Chatterjee (RA1511003040196)** , who carried out the project work under my supervision.

Signature of the Guide

Mr. N. Muthurasu
B.E., M.Tech,

Assistant Professor
Department of CSE
SRM IST
Vadapalani,Chennai-
600026

Signature of the HOD

Dr. S. Prasanna Devi
B.E., M.Tech, Ph.D.,
PGDHRM, PDF(IISc),
Head of the Department
Department of CSE
SRM IST
Vadapalani,Chennai-
600026

Acknowledgement

I place my gratitude with our respected chancellor of SRM Institute of Science and Technology, to have provided the required infrastructure, to hone my skills. I also thank my teachers for their guidance and teachings, without which I would not be able to learn what I know as efficiently as I do.

We extend this opportunity to extend our heartfelt gratitude to our respected Dean **Dr. K. Duraivelu** for his impeccable guidance.

We deeply express our sincere thanks to our Head of Department **Dr. S. Prasanna Devi**, for encouraging and allowing us to present the project on the topic, **News Analysis and Summarization Tool**, at our department premises for the partial fulfillment of the requirements leading to the award of B-Tech degree.

It is our privilege to express our sincerest regards to our project coordinator, **Mr. Muthurasu N.** for his valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of our project.

We take this opportunity to thank all our faculty members and management who have directly or indirectly helped our project. Last but not the least we express our thanks to our friends for their cooperation and support.

Abstract

In the modern day field of media, news is majorly affected by the social platforms such as Facebook, Twitter, etc. It becomes necessary to understand all the facts of an event, and to quantify, or relate them to understand it. Most humans do not have the dedication or time to study all the events at all the time, as time is always limited. However, an aggregation and analysis system on a computer could do it for them. A computer can form links and store endless amounts of data which would only be affected by the amount of storage space available and the storage algorithm being made use of. Apart from that advantage we have the advantage of having programs such as Hadoop and MongoDB to take care of a lot of work for us. This project has been worked upon with an idea in mind, to create a common platform for people to fetch aggregated data from multiple news agencies through an abstraction.

Contents

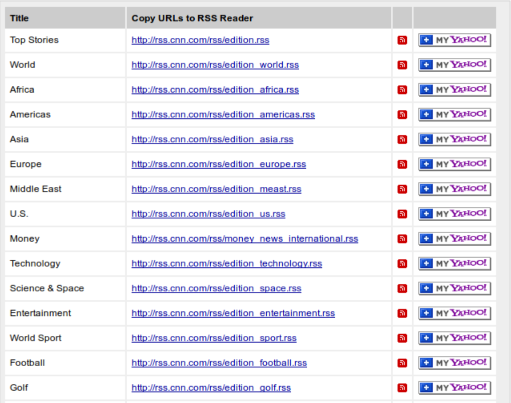
List of Figures

CHAPTER 1
Introduction

When dealing with news papers, it becomes a tedious chore to understand which one to follow and why. A lot of times, we see that in order for people to report an event, they sometimes have to withhold some little pieces only to reveal those at a later date. Apart from these, there is also the part where they need to be stored categorically to create a record of the modern history which keeps growing more and more opaque due to the complications everyday. So, how do we keep track of all facts while making sure we understand all related facets to an event?

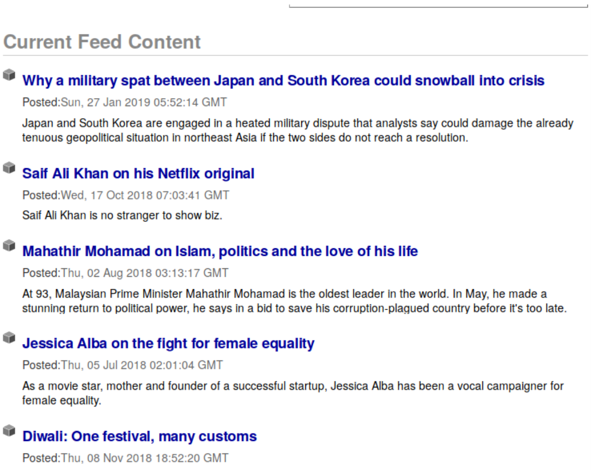
The solution is by relying on autonomous systems to work towards continuous content fetching, analyzing, aggregating and relating facts. But where can a person find the facts? It so happens that most major news channels offer a format known as RSS, to fetch the data. However, it is nigh impossible for us to keep track with the humongous number of news networks that exist within our country alone (the number being 1,05,443, of which, 13,661 are English Newspapers), let alone that of news channels all over the world. So in order to reduce the generation of noisy data, we reduce the scope to work with the most prominents ones.

The prominent news channels provide a certain utility known as an Atom feed, or RSS feed. These feeds have been shown in this section. These utilities are however under-utilized in most cases. Our work is to exploit this resource, made available from news networks, and manage it to maximize their usage.



Title	Copy URLs to RSS Reader		
Top Stories	http://rss.cnn.com/rss/edition.rss		MY YAHOO!
World	http://rss.cnn.com/rss/edition_world.rss		MY YAHOO!
Africa	http://rss.cnn.com/rss/edition_africa.rss		MY YAHOO!
Americas	http://rss.cnn.com/rss/edition_americas.rss		MY YAHOO!
Asia	http://rss.cnn.com/rss/edition_asia.rss		MY YAHOO!
Europe	http://rss.cnn.com/rss/edition_europe.rss		MY YAHOO!
Middle East	http://rss.cnn.com/rss/edition_meast.rss		MY YAHOO!
U.S.	http://rss.cnn.com/rss/edition_us.rss		MY YAHOO!
Money	http://rss.cnn.com/rss/money_news_international.rss		MY YAHOO!
Technology	http://rss.cnn.com/rss/edition_technology.rss		MY YAHOO!
Science & Space	http://rss.cnn.com/rss/edition_space.rss		MY YAHOO!
Entertainment	http://rss.cnn.com/rss/edition_entertainment.rss		MY YAHOO!
World Sport	http://rss.cnn.com/rss/edition_sport.rss		MY YAHOO!
Football	http://rss.cnn.com/rss/edition_football.rss		MY YAHOO!
Golf	http://rss.cnn.com/rss/edition_golf.rss		MY YAHOO!

Figure 1.1: CNN RSS Listings



Current Feed Content

Why a military spat between Japan and South Korea could snowball into crisis
Posted:Sun, 27 Jan 2019 05:52:14 GMT
Japan and South Korea are engaged in a heated military dispute that analysts say could damage the already tenuous geopolitical situation in northeast Asia if the two sides do not reach a resolution.

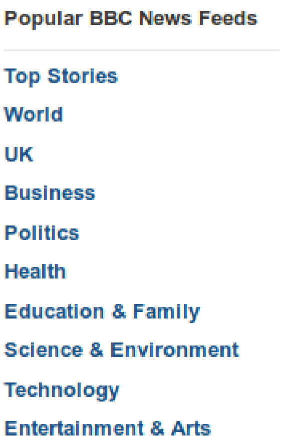
Saif Ali Khan on his Netflix original
Posted:Wed, 17 Oct 2018 07:03:41 GMT
Saif Ali Khan is no stranger to show biz.

Mahathir Mohamad on Islam, politics and the love of his life
Posted:Thu, 02 Aug 2018 03:13:17 GMT
At 93, Malaysian Prime Minister Mahathir Mohamad is the oldest leader in the world. In May, he made a stunning return to political power, he says in a bid to save his corruption-plagued country before it's too late.

Jessica Alba on the fight for female equality
Posted:Thu, 05 Jul 2018 02:01:04 GMT
As a movie star, mother and founder of a successful startup, Jessica Alba has been a vocal campaigner for female equality.

Diwali: One festival, many customs
Posted:Thu, 08 Nov 2018 18:52:20 GMT

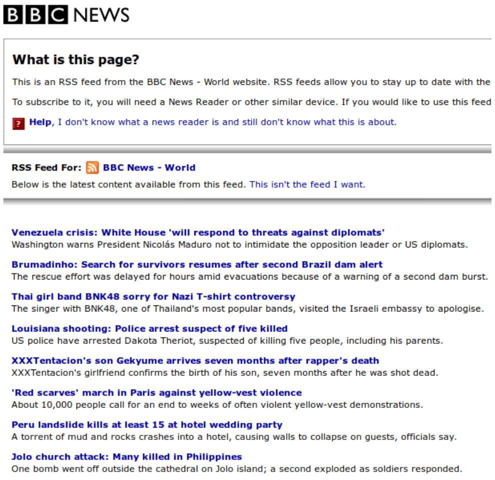
Figure 1.2: CNN RSS feed



Popular BBC News Feeds

Top Stories
World
UK
Business
Politics
Health
Education & Family
Science & Environment
Technology
Entertainment & Arts

Figure 1.3: BBC RSS Listings



BBC NEWS

What is this page?

This is an RSS feed from the BBC News - World website. RSS feeds allow you to stay up to date with the

To subscribe to it, you will need a News Reader or other similar device. If you would like to use this feed

Help, I don't know what a news reader is and still don't know what this is about.

RSS Feed For: **BBC News - World**
Below is the latest content available from this feed. This isn't the feed I want.

Venezuela crisis: White House 'will respond to threats against diplomats'
Washington warns President Nicolás Maduro not to intimidate the opposition leader or US diplomats.

Brumadinho: Search for survivors resumes after second Brazil dam alert
The rescue effort was delayed for hours amid evacuations because of a warning of a second dam burst.

Thai girl band BNK48 sorry for Nazi T-shirt controversy
The singer with BNK48, one of Thailand's most popular bands, visited the Israeli embassy to apologise.

Louisiana shooting: Police arrest suspect of five killed
US police have arrested Dakota Theriot, suspected of killing five people, including his parents.

XXXTentacion's son Gekyume arrives seven months after rapper's death
XXXTentacion's girlfriend confirms the birth of his son, seven months after he was shot dead.

'Red scarves' march in Paris against yellow-vest violence
About 10,000 people call for an end to weeks of often violent yellow-vest demonstrations.

Peru landslide kills at least 15 at hotel wedding party
A torrent of mud and rocks crashes into a hotel, causing walls to collapse on guests, officials say.

Jolo church attack: Many killed in Philippines
One bomb went off outside the cathedral on Jolo island; a second exploded as soldiers responded.

Figure 1.4: BBC RSS feed

CHAPTER 2
Project Work

Facts collection, aggregation and relation is everything this project is about. RSS feeds usually contain the shortened form of all the events, covering only the facts of the event.

2.1 The Concept

News articles contain well structured sentences that contain information in a human-readable language using the least amount of complicated words as can be managed.

To obtain the content, we make use of a Focused Incremental Web Crawler(A web crawler that extracts only specific content at a relatively shallow depth). This content is brought and stored for consideration in a pre-determined storage unit. This then moves on for analysis.

Facts aggregation is a lot about analysis of the text in the content, to find for three primary things, which are attributes, relations and involvement in events. These three are the specific information required. This involves the usage of NLP methodologies such as Named Entity Relations, Parts of Speech Tagging, Anaphora Resolution, and several others, which are to work together to reduce the amount of useless data to be stored about entities.

2.2 Problem Statement

Aggregation and analysis of news over network, from several channels to remove repetitions and increase efficient centralization of factual data of any event, location or an object.

2.3 Proposed Solution

The solution is to make use of an observer program to do it all, at regular intervals by utilizing a programming language, which in this particular case, for prototyping, would be through python.

2.4 Obstacles

There are three obstacles to this:

2.4.1 Relation Extraction

2.4.2 Anaphora Resolution

2.4.3 Metadata Repository

2.5 Advantages

2.6 Future Scope

2.7 Available tools

There are absolutely no tools available of this kind. There are several that forward the RSS feeds onto devices. But none that analyze to obtain all facts and aggregate it at a single point. A few of the more remarkable ones are specified here for your consideration.

2.7.1 News API

2.8 Tools Studied and Used

2.8.1 Python 3

2.8.2 nltk

2.8.3 Stanford CoreNLP

2.8.4 RSS

2.8.5 JSON

2.8.6 MongoDB

2.9 Theories

CHAPTER 3

Conclusion